



---

**To:** Joe Nett, Pete Wax, Aaron L, Kris Jensen, Tina Laidlaw, Janice Alers-Garcia, Jacques Oliver

---

**Cc:**

---

**From:** Kateri Salk, Mark Fernandez, Michael Paul

---

**Date:** 2020-12-08

---

**Subject:** NSTEPS North Dakota Lakes Progress Update

---

This memo covers the work conducted to date on the NSTEPS North Dakota Lakes project, spanning data compilation, data analysis planning, and classification analysis. Additional details are covered in:

- NSTEPS\_ND\_Lakes\_Data\_Compilation\_Report.docx
- NSTEPS\_ND\_Lakes\_DataAnalysisPlan\_v2\_20200924.docx

## 1. DATA COMPILATION

### 1.1. Data sources

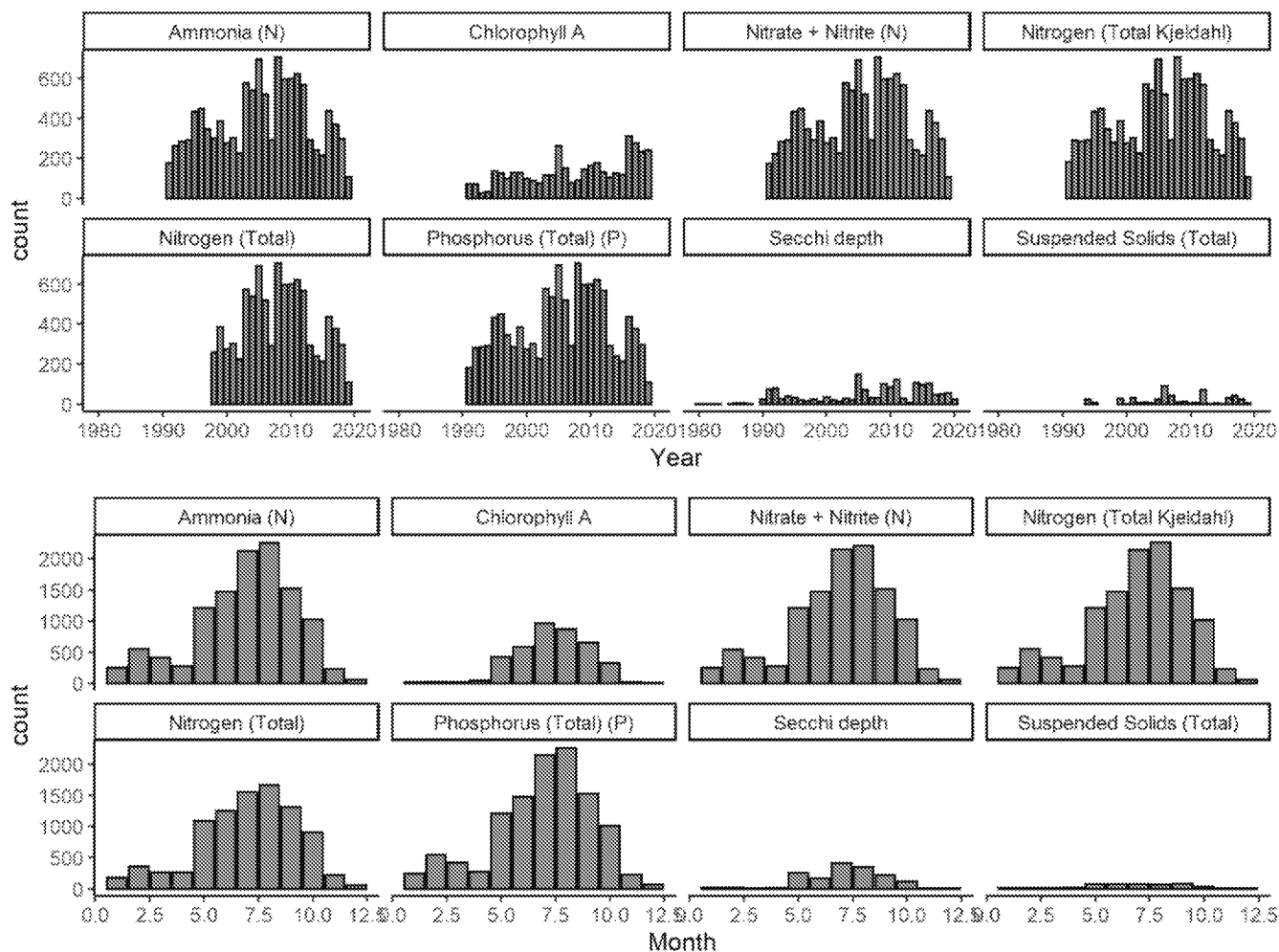
Four Excel files were supplied to NSTEPS. The “AllLakeNutrientData”, “AllProfileData”, and National Lakes Assessment (NLA) datasets are subsequently referred to as “Grab”, “Profile”, and “NLA.”

File Name	Description of Data	Format	Depths	Years
AllLakeNutrientData.xlsx	Nitrogen, phosphorus, bacteria, and TSS	Long	Varies	1991–2019
AllProfileData.xlsx	DO, pH, conductivity, and water temperature	Wide	Depth Profile	1980–2019
NLA2012_WQData_NSTEPS.xlsx	Chlorophyll a, nutrients, chemistry, and metals	Wide	Integrated Water Column	2012–2013
NLA2017_WQData_NSTEPS.xlsx	Chlorophyll a, nutrients, chemistry, and metals	Wide	Integrated Water Column	2017

### 1.2 Data processing

Data were processed in R (R Core Team 2019). Reported values of zero for typically log-transformed variables were replaced with half of the minimum nonzero value. Surface data ( $\leq 1$  m) were used in analyses unless otherwise stated. The NLA data were collected from a vertically integrated epilimnion and were considered surface.

Most of the data received was collected after 1990. Our intention was to use the most 10 recent years of data, per North Dakota Department of Water Quality (NDDWQ) assessment methodology. However, filtering for most recent 10 years severely reduced the number of available chlorophyll (chl) samples, so the full set of years was retained in the dataset pending further consideration and analysis. Samples were available from all months, but the data were filtered for the growing season, May-September, per the 2018 Integrated Report (IR). The IR also requires at least two samples per lake per season; no sample size limit was imposed on the dataset at this stage, but this requirement will be considered moving forward.



A Site ID common across the datasets is required to pair data spatially. The "AllLakesNutrientData" and "AllProfileData" both use a 6-digit numeric STORET ID. This ID was selected as the standard site ID. The NLA data used a different site ID format. To standardize the site IDs, all data was loaded into GIS, and NLA sites that were adjacent to a STORET ID were related as such. Non-adjacent NLA sites retained their NLA site ID. NLA site IDs varied across the study periods (2012 and 2017) for the same lake and therefore were standardized to the NLA 2017 format. There were 278 unique sites after data processing.

Lakes were identified using the North Dakota Lake Assessment Unit (AU) when available. Some NLA lakes were not able to be matched to a ND Lake AU. Those lakes were assigned an NHD Common ID (COMID) to uniquely identify them. There were 203 unique lakes after data processing.

### 1.3. Waterbodies missing information

Out of a total of 322 sites (some of which are located on the same lake), listed below is the number of sites that are missing information:

# sites missing data	Attribute
0	Station type (lake vs. reservoir, deepest site vs. other)
0	Latitude and Longitude
50	NHD ComID and ND LakeID
27	Level 3 Ecoregion
27	Elevation
27	HUC8
57	Watershed area and land use/land cover

The waterbodies missing information are:

- Devil's Lake
- Lake Sakakawea (some sites)

Only watershed information was missing for the following lakes:

- Heinrich-Martin
- Lutz Dam
- Indian Creek Dam
- Castle Rock Dam
- Harmony Lake
- Harmon Lake
- Long Lake
- East Park Lake
- Lightning Lake
- South Buffalo GAP
- Davis Dam
- White Earth Dam
- Custer Mine pond
- Carlson-Tande Dam
- Velva Sportsman's Dam
- Carbury Dam
- State Fair Pond
- Lake Alice – Lake Irvine

### 1.4. Addressing potential sources of bias

Some water bodies, specifically Lake Sakakawea, have a disproportionately large number of samples and sampling sites compared to the rest of the North Dakota lakes dataset. It was decided, in conversations with the

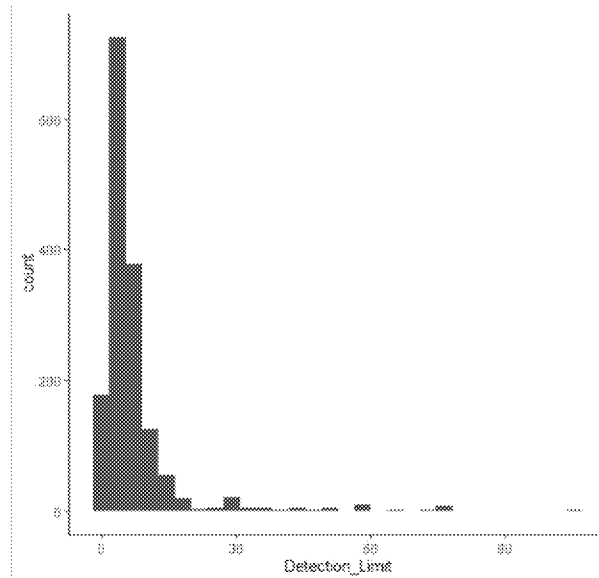
ND scientists, that Lake Sakakawea would not be part of this investigation, since it is the focus of other nutrient criteria efforts. However, it may be that Lake Sakakawea samples are helpful to the statewide modeling since they do represent a gradient of lake condition, albeit for a unique reservoir; therefore, it is important to consider retaining these samples. To avoid bias associated with drawing conclusions about statewide lakes from Lake Sakakawea, we will explore the distributions of variables of interest for statewide lakes vs. Lake Sakakawea as well as the regressions of bivariate relationships for these two groups as well. To the extent that samples from Lake Sakakawea are a representative distribution with respect to the rest of the statewide dataset, samples will be retained in the dataset. If Lake Sakakawea samples are found to introduce bias into the findings of analyses for statewide lakes, additional statistical approaches (including removing these samples) will be explored to reduce bias.

## 1.5. Treatment of non-detects

Values listed as below the detection limit for chlorophyll, ammonia, nitrate, total Kjeldahl nitrogen (N), total N, total phosphorus (P), and total suspended solids were imputed using robust regression on order statistics (ROS). The assumptions of log-normality for each variable were verified prior to imputation.

Variable	Non-detects	Total samples
Chl a	1,549	3,924
NH <sub>4</sub> <sup>+</sup>	4,356	11,393
NO <sub>2</sub> <sup>-</sup> + NO <sub>3</sub> <sup>-</sup>	4,741	11,344
TKN	44	11,407
TN	4	9,112
TP	679	11,405
TSS	194	475






Of a total of 3,924 total chlorophyll a (chl) samples, 2,375 were above the detection limit and 1,549 were below the detection limit. Of the samples below the detection limit, many had an unusually high detection limit. The overall range of reported detection limits spans typical ranges of chl concentrations found in lakes. Inclusion of these detection limits in the robust ROS imputation would thus lead to unrealistically high estimates of chl concentrations for some samples. The lowest chl concentrations reported in the dataset that were above the detection limit were between 2 and 3 µg /L, so it may be reasonable to assume that non-detects below 3 µg/L may be "true" nondetects and could be imputed reliably. In the portion of the dataset listed as below the detection limit, 829 had a detection limit ≤ 3 µg /L, whereas 720 had a detection limit > 3 µg /L. We moved forward with imputing non-detect values via robust ROS for the former group, and we removed the samples from the dataset for the latter group.

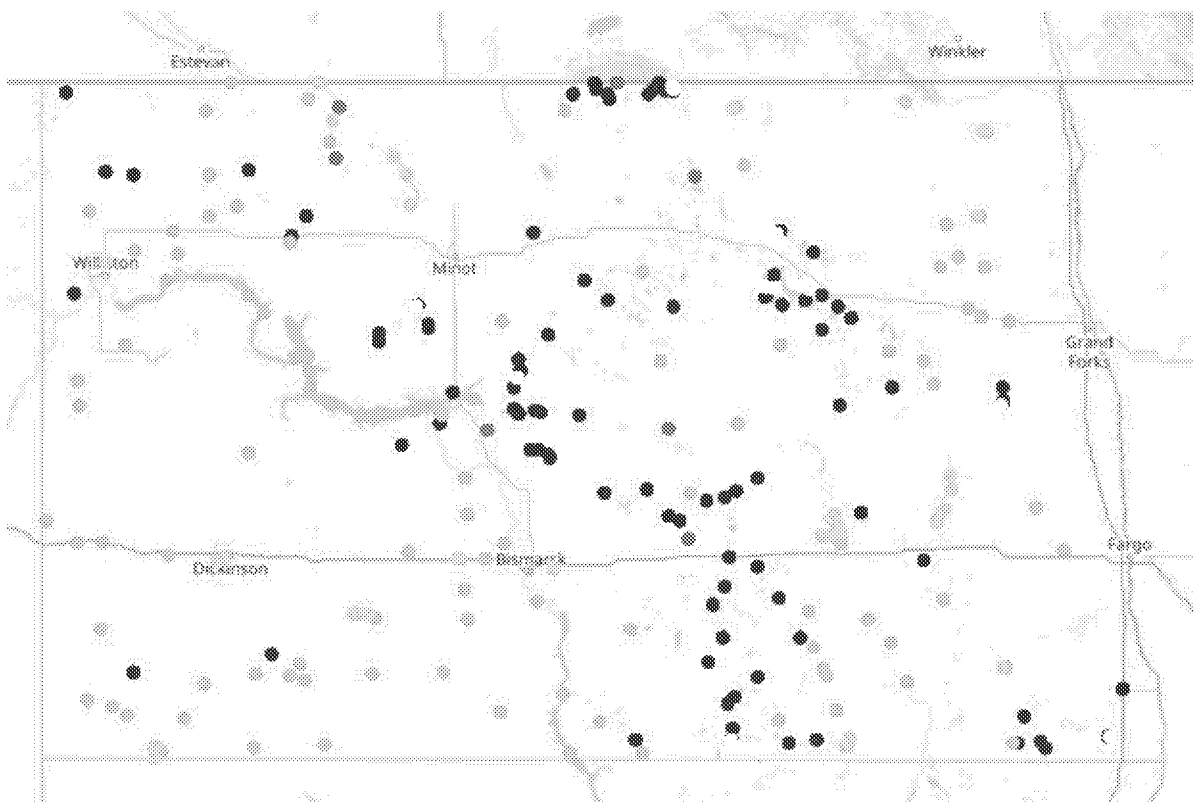


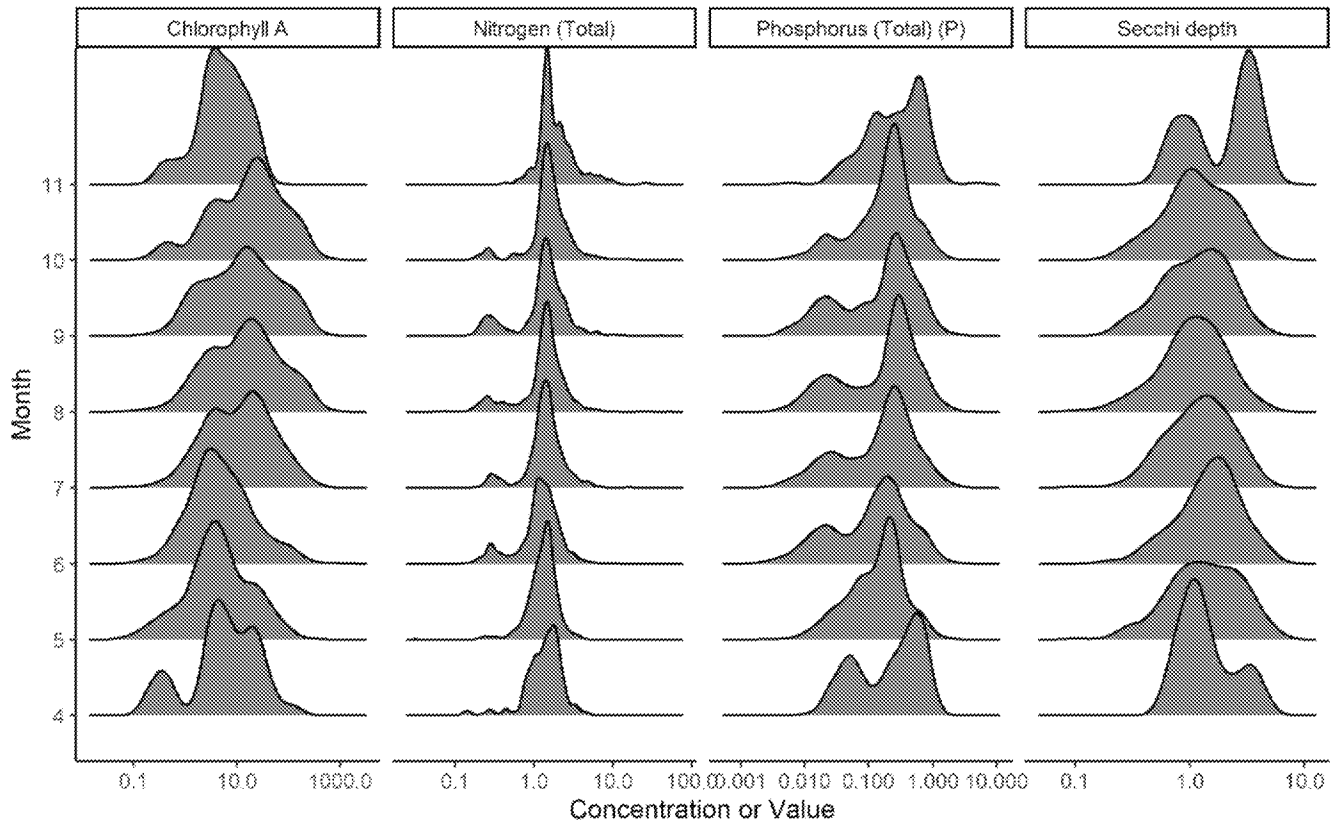
Distribution of reported detection limits for the chl dataset, in µg/L. Detection limits were only reported for samples whose concentration was below the detection limit.

## 1.6. Data exploration

Data were explored spatially and temporally. Examples of visualizations are displayed below.

Station Type    Lake    Lake (Deepest)    Lake (Other)    Reservoir (Deepest)    Reservoir (Other)





## 2. DATA ANALYSIS PLANNING

### 2.1 Classification

Several variables may help to split lakes into groupings based on their distributions of specific variables as well as bivariate relationships among variables. Classification approaches seek to find groupings of lakes that exhibit differential behavior with respect to responses to nutrients.

Relevant variables for classification analysis include:

- Ecoregion
- Geology
- Lake vs. reservoir (or Station Type)
- Depth
- Lake Area
- Elevation
- N:P ratio
- Non-algal turbidity

Additional variables that may be useful, but we do not have in the dataset at this time, are:

- Use class
- Thermal regime
- Residence time

We propose to implement the following explicit classification analyses:

1. Explore differences in the distributions of N, P, and chl among proposed classes
2. Conduct model-based recursive partitioning (MBRP) of chl ~ N and chl ~ P using the variables above as candidate classifying variables. MBRP is like classification and regression trees but instead of differences in the distributions of responses (e.g., deviance in chl among classes), one looks for differences in functional responses on either side of the split in the data. The approach seeks functional differences in the chl ~ nutrient relationships as the classifier varies.

## 2.2. Exploratory Data Analysis

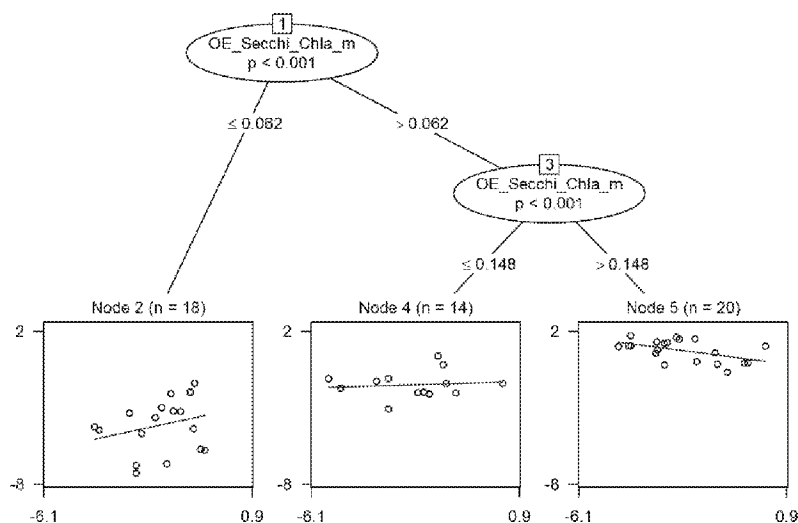
We will first explore the distributions and sample sizes of variables of interest (chlorophyll, nutrients, clarity, phytoplankton, oxygen, pH) among sample classes. Aggregation approaches (individual observations, seasonal averages, etc.) will be explored as well. Basic visualizations of univariate and bivariate relationships will be produced.

## 2.3. Stressor-Response Analysis

In order to identify quantitative nutrient and chlorophyll thresholds associated with adverse response conditions in lakes, we will conduct stressor-response analysis among lake classes. Predictor variables (TN, TP, chlorophyll) and response variables (chlorophyll, phytoplankton biomass and assemblage, cyanobacteria abundance, pH, dissolved oxygen) will first be modeled in bivariate space to assess potential linear and nonlinear responses. Change point analysis will be deployed for relationships where thresholds may be at play. Then, potential confounding factors including land use, non-algal turbidity, N:P ratios, and thermal regime will be added to models in generalized linear models and/or hierarchical models.

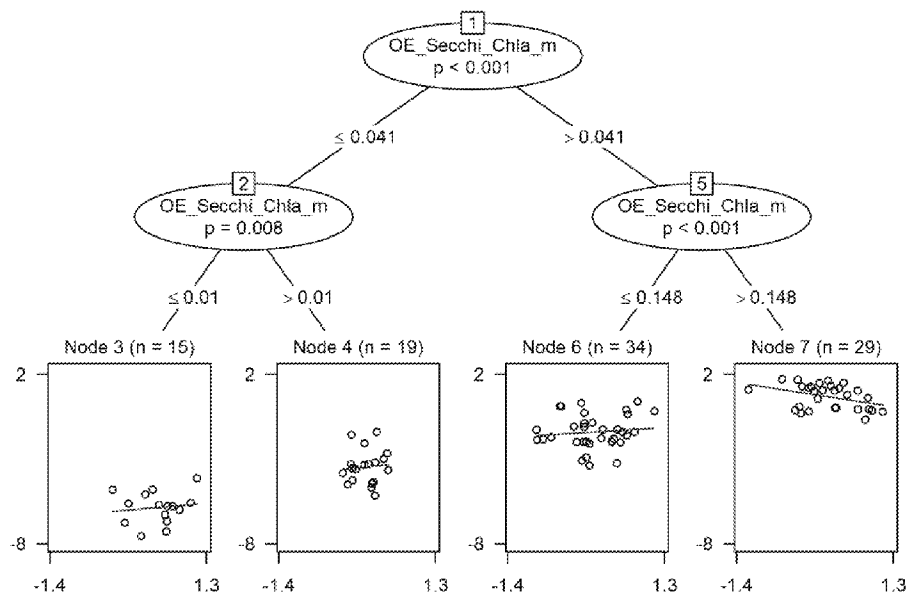
## 3. CLASSIFICATION ANALYSIS

As described above, MBRP was used to identify statistical splits in the relationship between chl and nutrients (TN and TP) based on potential classifier variables. The first round of MBRP analysis included only data from the last 10 years. However, the splitting resulted in sample sizes that were too small to discern meaningful relationships among variables. An example of chl ~ TP, with Secchi O/E values as the splitting variable, is shown below.

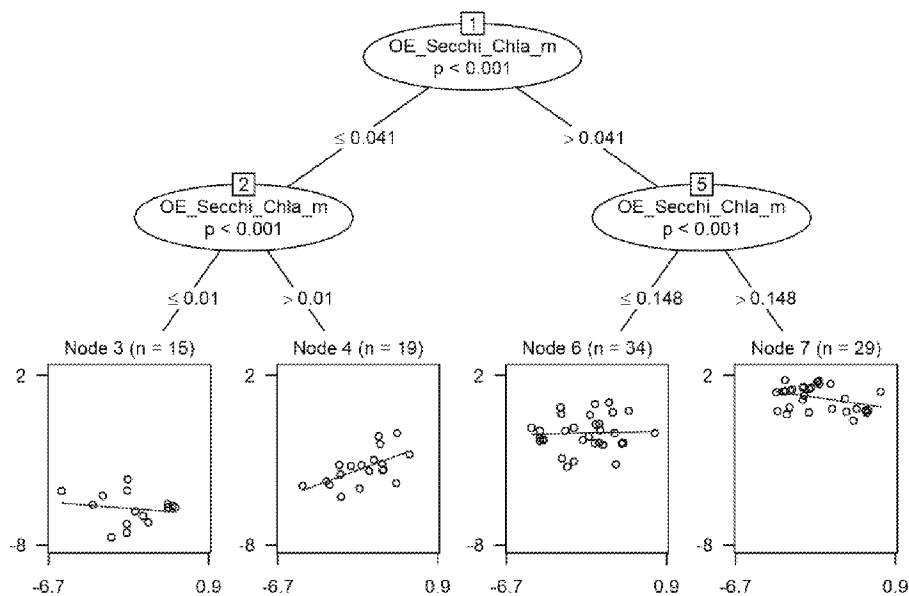


When all years of data were included and all potential classifier variables were included, Secchi O/E was selected as the most important variable.

Chl ~ TN:

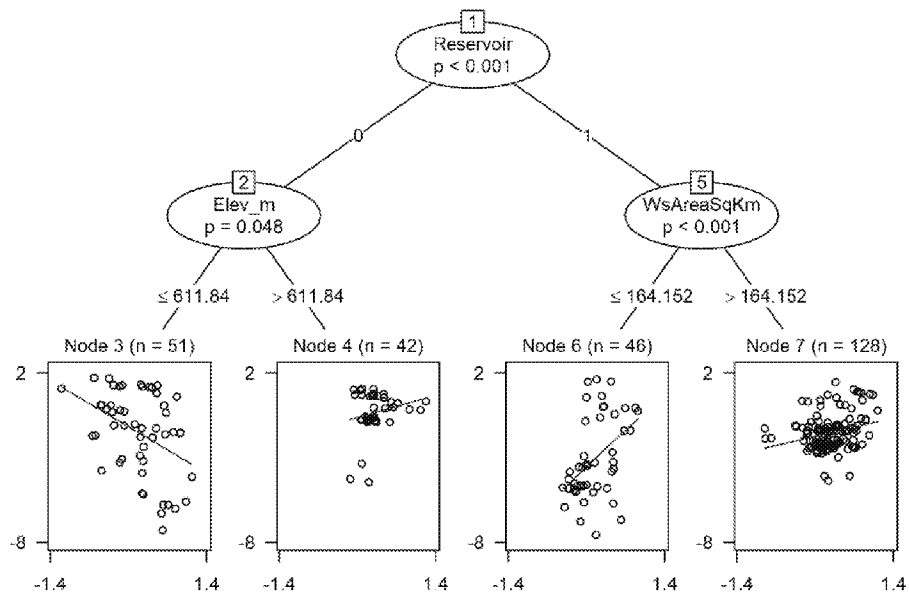


Chl ~ TP:

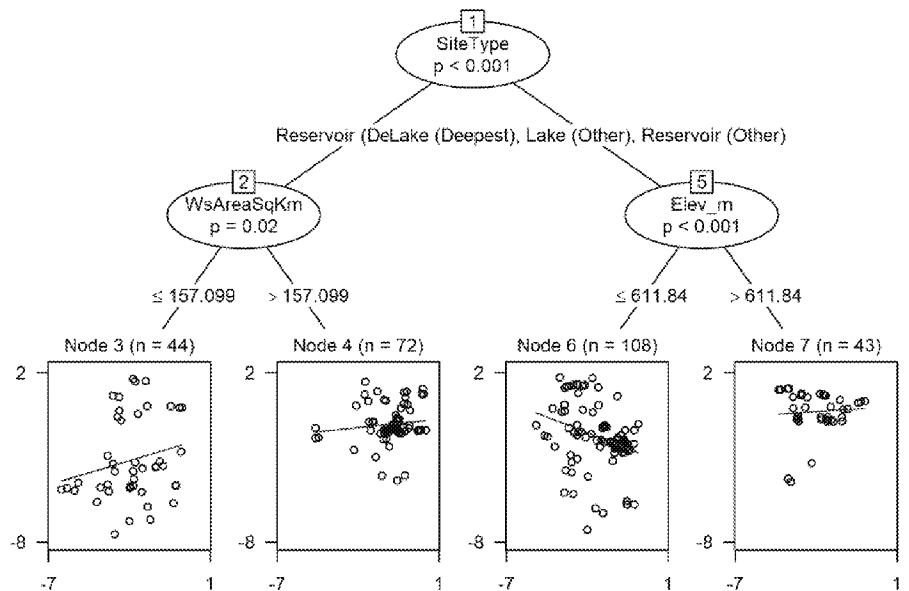


However, when Secchi O/E was included as a predictor variable, the sample sizes in the dataset were still limited. When clarity variables were left out of the analysis, sample sizes were larger and other variables emerged as splitting variables including the site or waterbody type, elevation, and watershed area.

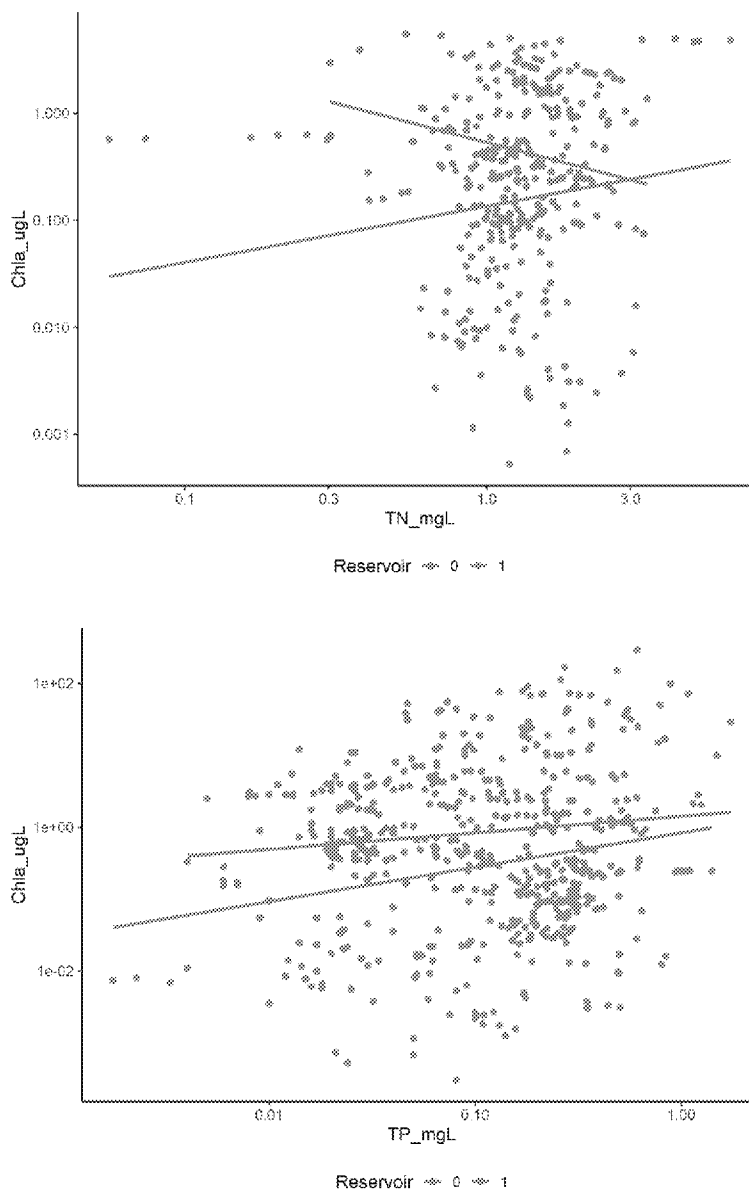
Chl ~ TN:



Chl ~ TP:



When the first splitting variable, the waterbody type (lake or reservoir) was analyzed as the sole classification variable, the relationships of chl and nutrients are as follows:



Waterbody type was a significant predictor of chl, and there were significantly different slopes in the chl~TN relationship between the waterbody types ( $df = 337$ ,  $p < 0.01$ ). Waterbody type was a significant predictor of chl, but the slopes in the chl~TP relationship were not significantly different between waterbody types ( $df = 565$ ,  $p = 0.19$ ).